Network Inversion for Uncertainty-Aware Out-of-Distribution Detection

Out-of-distribution (OOD) detection and uncertainty estimation (UE) are critical components for building safe machine learning systems, however the two problems have, until recently, separately been addressed. A key limitation of modern neural networks is their tendency to produce overconfident predictions even on inputs that lie far outside the training distribution. This makes it crucial to develop models capable of both OOD detection—the ability to identify inputs that fall outside the training distribution—and UE—the ability to quantify confidence in predictions to ensure safe decision-making under distributional shift.

In this work, we propose a novel framework that leverages network inversion [1], not only to detect OOD inputs but also to estimate prediction uncertainty, unifying the two objectives in a single training procedure. For a standard n-class classification task, we extend the classifier to an n+1 class model by introducing "garbage" class, initially populated with random gaussian noise to represent outlier inputs. After each training epoch, we use network inversion to reconstruct input images corresponding to all output classes. These inverted images initially appear noisy and incoherent with highly uncertain labels; hence excluded to the garbage class for retraining the classifier. This cycle of training, inversion, and exclusion continues till the inverted samples begin to resemble the in-distribution data more closely as shown in the figure with a significant drop in the uncertainty, suggesting that the classifier has learned to carve out meaningful decision boundaries while sanitising the class manifolds by pushing OOD content into the garbage class.

In each subsequent epoch the classifier is trained using a weighted cross-entropy loss to account for the class imbalance introduced by addition of garbage samples. Unlike prior approaches, our method requires no external OOD datasets or post-hoc calibration, offering a simple and interpretable solution to ensure robustness in classification under distributional shift. During inference, this training scheme enables the model to effectively detect and reject OOD samples by classifying them into the garbage class. The confidence scores corresponding to class predictions can be used to assess the model's uncertainty. Low confidence on in-distribution predictions indicates ambiguous or uncertain inputs, while high confidence in the garbage class suggests a strong belief that the input is OOD. We quantify uncertainty using the confidence values across all n+1 output classes by capturing how sharply peaked or spread out the model's predictive distribution $\mathbf{p}$ is, given by

$$\mathrm{UE}(\mathbf{p}) = 1 - \frac{\sum_{i=1}^{n+1}\left(p_i - \frac{1}{n+1}\right)^2}{\sum_{i=1}^{n+1}\left(\delta_{i,k} - \frac{1}{n+1}\right)^2}$$

Train \ Test	MNIST	FMNIST	SVHN	CIFAR-10
MNIST	99.1	89.5	99.1	99.4
FMNIST	85.2	92.6	96.3	95.7
SVHN	93.6	94.9	89.4	87.6
CIFAR-10	97.8	95.7	88.2	85.5

where $k = \arg\max_i p_i$ and $\delta_{i,k}$ is the *Kronecker* delta. The resulting score ranges from 0(min uncertainty) to 1(max uncertainty), providing an interpretable measure of uncertainty.

We evaluate the effectiveness of our approach to uncertainty-aware out-of-distribution detection across four benchmark image classification datasets following a one-vs-rest evaluation strategy: the model is trained exclusively on one dataset and evaluated on the remaining three as OOD sources. The table above presents the accuracy for uncertainty-aware OOD detection across all pairs of datasets. Each row corresponds to a model trained on one of the datasets and diagonal entries represent the in-distribution (ID) performance measured on the standard test set of the training dataset. Off-diagonal entries indicate OOD detection performance, where the accuracy represents how well the model distinguishes out-of-distribution samples by correctly classifying them into the garbage class.

High values across both diagonal and off-diagonal entries demonstrate that the model maintains strong classification performance on ID data while reliably identifying OOD inputs. While the majority of OOD samples are assigned to the garbage class, a few can still be misclassified into in-distribution classes. However, on average, we observe that the least confidently classified in-distribution sample is still more confidently classified compared to the most confidently misclassified out-of-distribution sample, suggesting the existence of a clear threshold for further improvement. Future work can also consider the use n garbage classes—one for each of the in-distribution classes—for fine-grained separation of OOD samples and weighted individual OOD sample contribution to the loss while retraining the classifier based on uncertainty.

[1] Pirzada Suhail and Amit Sethi. Network Inversion of Convolutional Neural Nets. AAAI 2025